OC3D: Weakly Supervised Outdoor 3D Object Detection with Only Coarse Click Annotation

Hongwei Lin 23020241154352[1], **Longqing Chen 31520241154496**[1], **Yaoxin Chen 36920241153200**[2]

[1] Informatics Class
[2] AI Class

Abstract

LiDAR-based outdoor 3D object detection has received widespread attention. However, training 3D detectors from the LiDAR point cloud typically relies on expensive bounding box annotations. This paper presents **OC3D**, an innovative weakly supervised method requiring only coarse clicks on the bird's eye view of the 3D point cloud. A key challenge here is the absence of complete geometric descriptions of the target objects from such simple click annotations. To address this problem, our proposed OC3D adopts a two-stage strategy. In the first stage, we initially design a novel dynamic and static classification strategy and then propose the **Click2Box** and **Click2Mask** modules to generate box-level and mask-level pseudo-labels for static and dynamic instances, respectively. In the second stage, we design a **Mask2Box** module, leveraging the learning capabilities of neural networks to update mask-level pseudo-labels, which contain less information, to box-level pseudo-labels. Experimental results on the widely used KITTI and nuScenes datasets demonstrate that our OC3D with only coarse clicks achieves state-of-the-art performance compared to weakly-supervised 3D detection methods. Combining OC3D with a missing click mining strategy, we propose an OC3D++ pipeline, which requires only 0.2% annotation cost in the KITTI dataset to achieve performance comparable to fully supervised methods. The code will be made publicly available.

Introduction

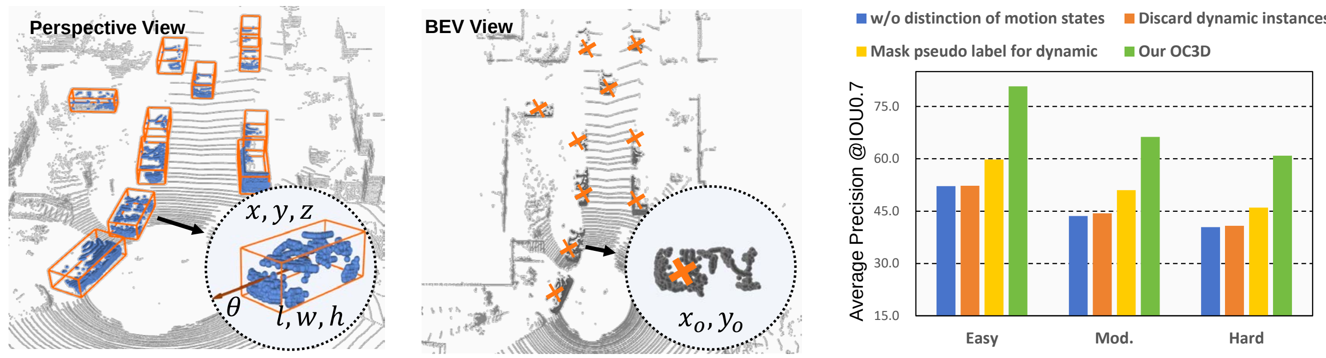
In recent years, notable progress has been made in LiDAR-based 3D object detection research (Wu et al. 2023). Despite these advancements, the need for precise bounding box supervision remains a major challenge due to its time-consuming and labor-intensive nature. For instance, the KITTI dataset (Geiger, Lenz, and Urtasun 2012) contains 3,712 training scenes with over 15,000 vehicle instances, where manual annotation of a single instance can take roughly 114 seconds (Meng et al. 2021). The extensive labeling efforts required escalate dramatically when scaling detectors to larger-scale datasets (Sun et al. 2020; Caesar et al. 2020), thereby hindering further research in a fully supervised manner.

To alleviate the annotation burden, recent studies have explored alternatives that require *fewer* annotated frames or instances to train high-performing 3D object detectors. Specifically, *semi-supervised methods* (Wang et al. 2021) leverage a subset of the annotated frames, while *sparsely-supervised* methods (Xia et al. 2023; Liu et al. 2022; Xia et al. 2024) rely on only one bounding box annotation per frame during training. While these approaches have significantly lowered annotation costs, annotating 6DoF bounding boxes for every scene remains time-consuming.

To provide a faster, albeit less precise, method of human supervision, WS3D (Meng et al. 2021) and ViT-WSS3D (Zhang et al. 2023) propose a *mixed supervision* strategy that replaces some box annotations with the **center-click** annotations. In this approach, annotators click the center of objects on the Bird's Eye View (BEV) to generate center-level labels, reducing labeling time per instance to approximately 2.5 seconds, which is 50 times faster than traditional bounding box labeling.

However, relying on a single center point has significant limitations: (1) it requires annotators to precisely indicate the *center* position and (2) it fails to accurately represent the *shape* and *scale* of objects, especially in sparse point clouds where objects may be partially observed. These challenges become even more severe for *moving objects*, where motion across frames further complicates the estimation of an accurate bounding box from clicks. As a result, prior works have struggled to scale up the use of click supervision, combined with traditional box annotation for mixed supervision, which diminishes the overall effectiveness.

In this paper, we introduce a novel pure click-supervised approach for 3D object detection (**OC3D**), which employs temporal cues to distinguish between static and moving instances and progressively recover their box supervision from coarse clicks (Fig. 1 (b)) for detector training. This strategy offers greater flexibility than center-clicked approaches by accommodating inaccurate clicks through shift tolerance, which can be corrected via multiple-frame consistency. Our framework comprises three key designs: (1) To efficiently classify the motion state of provided clicks, we analyze the point density at the clicked location from a statistical perspective, noting that *vary* point density over a short period for moving objects while *stable* density for static instances. (2) For static objects, we implement a **Click2Box** strategy that aggregates the neighboring points across multiple frames to reconstruct the object's structure and lift click-level supervision to a precise 3D box. (3) For moving objects, which are more challenging to regress, we generate



(a) Fully-supervised Annotation **(b) Coarse-click Annotation** **(c) Performance Comparison of Different Strategies**

Figure 1: (a) 6DoF box annotation; (b) Unlike costly box annotations, coarse-click alternatives require only quick clicks on objects in the 2D BEV plane, yet offer limited supervision; (c) An empirical study on maximizing the utility of weak click labels reveals significant performance gains when separating static and dynamic objects (yellow bars) compared to directly estimating 3D bounding boxes from clicks (blue bars) or discarding pseudo labels for dynamic objects (orange bars). The proposed OC3D (green bars) further refines label quality via Mask2Box conversion, substantially enhancing detection performance by 17%.

mask-level pseudo labels (**Click2Mask**) to assign the clustered points with labels, enabling models to learn to predict the object's location. To compensate for the lack of supervision in predicting shapes and scales for moving objects, we leverage their geometric similarity as a bridge, further refining the 3D box through a **Mask2Box** process. We evaluated OC3D on the widely adopted KITTI and nuScenes datasets. Remarkably, OC3D achieves competitive performance with weakly-supervised baselines that rely on accurate box annotations.

In summary, our contributions are:

- We propose the first method of only click annotated 3D object detection from point cloud (OC3D), which solely relies on coarse clicks on the BEV maps. This approach dramatically reduces the annotation cost of 3D object detection tasks to $1\% \sim 0.2\%$.
- We design the **Click2Box** and **Click2Mask** modules according to the motion attributes of instances, inferring accurate mixed supervision information from the click annotations.
- We design a **Mask2Box** module that upgrades the less informative mask-level pseudo labels to box-level pseudo labels, which compensates for the loss of object shape and scale information in mask-level supervision.

Related Work

LIDAR-based 3D Object Detection. In recent years, fully-supervised 3D object detection has been widely studied. The early methods (Lang et al. 2019; Yan, Mao, and Li 2018; Yin, Zhou, and Krähenbühl 2021) utilized an end-to-end one-stage object detection strategy, predicting detection boxes directly from point clouds. Subsequently, the two-stage methods (Shi, Wang, and Li 2019; Deng et al. 2021; Shi et al. 2023; Wu et al. 2022; Chen et al. 2023; Wang et al. 2024) introduced an additional proposal box refinement stage, which improves detection performance by refining regions of interest. Despite achieving excellent performance, all these methods require costly box annotations, the generation of which is time-consuming and labor-intensive.

Label-efficient 3D Object Detection. Research on reducing annotation costs in 3D object detection tasks has received widespread attention. Around semi-supervised methods (Wang et al. 2021), they selected only a small number of fully annotated frames as labeled data, using the remaining frames as unlabeled data. These methods used teacher-student networks for distillation learning to mine and generate pseudo-labels. Sparsely-supervised methods (Xia et al. 2023; Liu et al. 2022) adopted a sparsely annotated strategy, retaining only one complete bounding box label for each selected frame. They utilized specially designed unlabeled object mining modules to discover potential pseudo-labels. Although these strategies have significantly reduced the dependence on 3D boxes, it is still not possible to completely abandon laborious box-level annotations.

Proposed Solution

Overview As shown in Fig. 2, given the coarse nature of click supervision, our approach proceeds by first (1) generating mixed-granularity supervision for both static and moving objects using the Click2Box and Click2Mask strategies, followed by (2) applying Mask2Box-enhanced self-training to refine and improve the quality of generated supervision. The specific steps involved in this process are elaborated as follows.

Mixed Pseudo-Label Generation

To accurately estimate 3D bounding boxes from click annotations as pseudo labels, we aim to leverage temporal cues to enrich sparse point annotations by aggregating registered points across consecutive frames. For static instances, the aggregation of dense points facilitates the complete capture of spatial geometric information of the instance. However, for dynamic objects, the dense point aggregation often fails to capture the full geometric structure, leading to low-quality

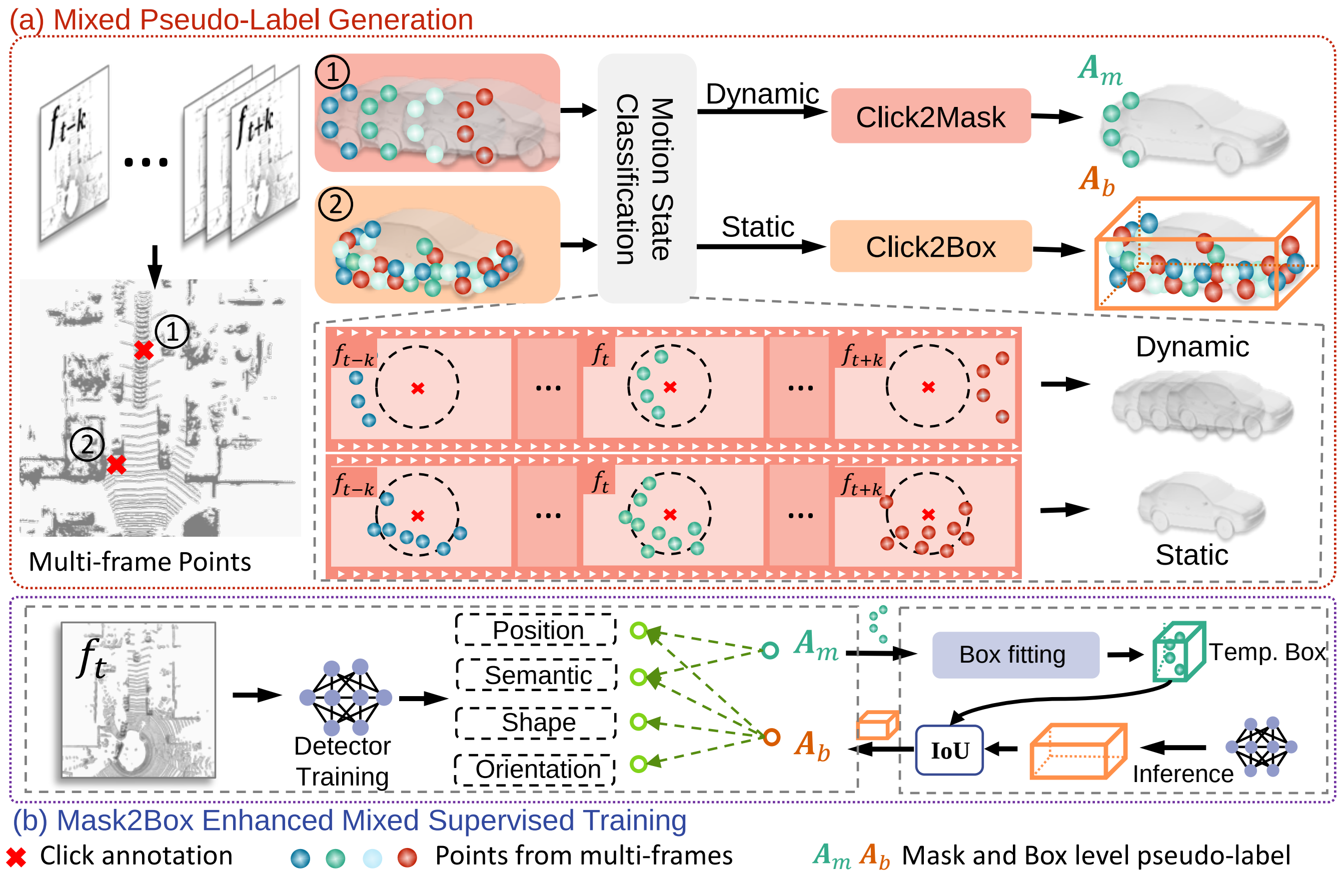


Figure 2: The pipeline of the proposed OC3D. (a) Initially, a novel motion state classification strategy is introduced, followed by the generation of mask-level pseudo-label $\mathbf{A}_m$ and box-level pseudo-label $\mathbf{A}_b$, utilizing the Click2Mask and Click2Box modules, respectively. (b) With the mixed pseudo-labels generated by stage (a), train the detector and then update the mask-level supervision to box-level based on high-confidence predictions of the trained detector.

pseudo-labels. This challenge motivates us to first classify the motion status of instances corresponding to each annotation click:

Motion State Classification for Clicked-instance. We observe the duration of local point distribution at clicked positions during a long sequence traversal. Specifically, for static instances, the local point cloud at the clicked position exhibits a continuous distribution, whereas, for dynamic instances, the local point cloud at the clicked position is transient throughout the traversal. Motivated by this, we utilize the persistence of points at local positions within the long sequence for dynamic and static classification.

For each click annotation $c_o = (x_o, y_o)$ at the t-th frame, we gather adjacent frames $\mathcal{F} = \{f_{t-k}, ..., f_t, ..., f_{t+k}\}$ within a local time window k, followed by ground removal (Himmelsbach, Hundelshausen, and Wuensche 2010) as a preprocessing step. Our primary focus is on the BEV points in $\mathcal{F}$, denoted as $\{\mathbf{P}_t^{\mathrm{BEV}} \in \mathbb{R}^{N\times 2}\}_{t\in[t-k,t+k]}$, where N indicates the number of BEV points in each frame. To determine the persistence of the clicked position (x_o, y_o), we search for its neighboring BEV points within a radius r, resulting in the collection $\{\mathbf{N}_t\}_{t\in[t-k,t+k]}$, with each time step having a cardinality of N_t:

$$\mathbf{N}_t = \left\{p_i \in \mathbf{P}_t^{\mathrm{BEV}} \mid \|p_i - c_o\|_2 \le r\right\}, N_t = |\mathbf{N}_t|. \quad (1)$$

To better tally the duration for which points are continuously present near the clicked location, we construct the function $g(t)$, and perform a differentiation operation on $g(t)$:

$$g(t) = \begin{cases} 0 & \text{if } N_t = 0; \\ 1 & \text{otherwise.} \end{cases} \quad (2)$$

$$\Delta g(t) = g(t+1) - g(t). \quad (3)$$

In adjacent T frames ($T = 2k+1$), the duration of the neighborhood points of the click position can be calculated based on the index of the specific value of $\Delta g(t)$. That is, $\Delta g(t) = 1$ indicates the point begins to appear, $\Delta g(t) = -1$ indicates the point disappears, and marking the time difference between the last appearance and the next disappearance of the point on the click annotation frame is the duration time Δt. Subsequently, the motion state of the clicked-instance is determined based on the proportion of Δt that occupies adjacent T frames.

$$\begin{cases} \texttt{static} & \text{if } \frac{\Delta t}{T} > \tau; \\ \texttt{dynamic} & \text{otherwise.} \end{cases} \tag{4}$$

where τ is the duration threshold. If $\frac{\Delta t}{T}$ exceeds the threshold, it indicates that the local point cloud around the click has a longer duration and is considered a static instance. Conversely, if it does not exceed the threshold, it is considered a dynamic instance.

Click2Box. For static objects, dense points express complete geometric structures, which supports the fitting of high-quality bounding box pseudo-labels from the point cloud distribution. Motivated by this observation, we concatenate the neighboring points of multiple frames $\{\mathbf{N}_t\}_{t\in[t-k,t+k]}$ to obtain local dense points $\mathbf{D}_t$ for the time step t. We perform DBSCAN (Ester et al. 1996) clustering algorithm on $\mathbf{D}_t$ to generate several discrete point clusters. We retain the cluster of points whose center is closest to the clicked position and consider the points in this cluster as the foreground points of the clicked instance. Finally, we perform a bounding box fitting algorithm (Zhang et al. 2017) on the foreground points to generate a box-level pseudo-label. We utilize Click2Box to infer box-level pseudo-labels $\mathbf{A}_b$ from the click annotations of all static instances.

Click2Mask. For dynamic objects, we opt to leverage only the single-frame point cloud $\mathbf{P}_t$ due to the long-tail distribution observed in aggregated points resulting from motion differences (Chen et al. 2022). Although the instance shape and scale cannot be revealed by click-level labels, the foreground points from click-annotated frame can still provide reliable *semantic* information and coarse *location* information. Consequently, instead of generating box-level pseudo labels for dynamic objects, we produce mask-level pseudolabels by extracting the foreground points in $\mathbf{P}_t$. To identify these foreground points, we employ the DBSCAN clustering algorithm on raw point clouds, and select the cluster with the center closest to the click location a mask-level pseudo label. The resulting mask-level pseudo labels denoted as $\mathbf{A}_m$, are derived from the click annotations of all dynamic instances.

Mask2Box Enhanced Mixed Supervised Training

In contrast to traditional 3D object detectors that are solely reliant on box supervision, our approach delves into weakly supervised detectors with mixed supervision. Inspired by MixSup(Yang, Fan, and Zhang 2024), we re-engineered the strategy for supervision allocation. However, the information about mask-level supervision is limited, and it is challenging to achieve optimal detector performance. To address this issue, combining with an iterative training, we introduce a mask2box enhanced training strategy, which leverages the high-confidence outputs from the last iterative detector to refine mask-level pseudo-labels into more accurate bounding box-level pseudo-labels, thereby improving the overall precision of the detection process.

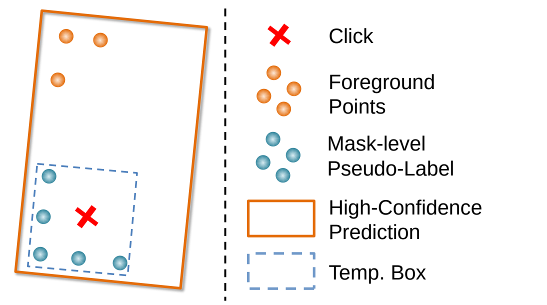


Figure 3: An illustration of the upgrading strategy for mask-level pseudo labels.

Mixed Supervised Training. Referencing (Yang, Fan, and Zhang 2024), we redesigned the supervision assignment strategy. For all box-level pseudo-labels $\mathbf{A}_b$ of size A_b, we train the detection network to focus on its position, shape, orientation, and semantics. For mask-level pseudo-labels $\mathbf{A}_m$ of size A_m, the detector focuses solely on semantics and the center of masks. Therefore, the loss function for OC3D can be formulated as:

$$\mathfrak{L} = \frac{1}{A_b}\sum_{A_b}\mathfrak{L}_{\text{reg}} + \frac{1}{A_b + A_m}\sum_{A_b+A_m}\mathfrak{L}_{\text{cls}} + \lambda\frac{1}{A_m}\sum_{A_m}\mathfrak{L}_{\text{pos}}. \tag{5}$$

$\mathfrak{L}_{\text{reg}}$ and $\mathfrak{L}_{\text{cls}}$ are commonly used regression and classification losses in 3D object detection. $\mathfrak{L}_{\text{pos}}$ is the part that decouples the central position from $\mathfrak{L}_{\text{reg}}$. Since mask centers are not accurate instance centers, we set the hyper-parameter λ to reduce the weight of this part of the loss.

Mask2Box Enhanced Training. Mask-level pseudo-labels provide only semantic and coarse localization information, lacking a description of the instance shape. To compensate for the information lost in mask-level pseudo-labels, we use high-confidence bounding-box predictions to upgrade mask-level pseudo-labels. As shown in Fig. 3, based on mask-level pseudo-labels, we use the box fitting algorithm (Zhang et al. 2017) to generate a temporary bounding box (Temp. Box). Then, we calculate the IoU of the temporary box with all high-confidence predictions and return the prediction box with the highest IoU:

$$\text{Proposal} = \underset{\text{IoU}}{\arg\max}\,(\text{IoU}_i = \frac{\text{Temp. Box} \cap l_i}{\text{Temp. Box} \cup l_i}). \tag{6}$$

where l_i originates from all the high-confidence predictions. If the point-level pseudo-label is within the proposal, then the proposal is used to replace the point-level pseudo-label. Conversely, the point-level pseudo-label is retained and will be awaited for the next upgrade opportunity.

Experiments

Datasets and Metrics. We evaluated OC3D on the widely adopted KITTI (Geiger, Lenz, and Urtasun 2012) and nuScenes (Caesar et al. 2020) datasets. Remarkably, OC3D achieves competitive performance with weakly-supervised baselines that rely on accurate box annotations. For evaluation, we adopt mAP and nuScenes detection score (NDS) as the main metrics.

Method	Venue	Annotations		3D-Detection			BEV-Detection		
		Type	Cost	Easy	Mod	Hard	Easy	Mod	Hard
Voxel-RCNN (Deng et al. 2021)	AAAI2021	Fully Supervised	100%	98.7	94.9	94.5	98.7	94.9	94.6
WS3D (Meng et al. 2021)	PAMI2021	Boxes + Center Clicks	3%	96.3	89.4	88.9	96.4	89.3	88.9
MixSup (Yang et al. 2024)	ICLR2024			94.9	92.7	90.0	94.9	93.0	90.4
SS3D (Liu et al. 2022)	CVPR2022	Sparse Boxes	2%	98.3	89.2	88.3	-	-	-
CoIn (Xia et al. 2023)	ICCV2023			96.3	86.7	74.4	96.3	88.4	75.6
CoIn++ (Xia et al. 2023)	ICCV2023			99.3	92.7	88.8	-	-	-
HINTED (Xia et al. 2024)	CVPR2024			98.5	91.6	90.3	98.4	92.9	90.6
OC3D++	-	Sparse Coarse Clicks	**0.2%**	96.4	91.6	84.6	96.5	92.0	84.9
OC3D	-	Coarse Clicks	**1%**	96.6	92.5	91.5	96.7	94.1	92.0

Table 1: Experimental results on KITTI dataset compared with recent state-of-the-art label-efficient methods. We report results of 'car' with 40 recall positions, below the 0.5 IoU thresholds. 'Boxes + Center-Clicks' denotes that in some scenarios, bounding box annotations are retained, while in the rest, only the central position annotations are preserved. 'Sparse Boxes' indicates that only one bounding box annotation is retained per ten scenes.

Method	Annotation	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bike	Ped.	T.C.
CenterPoint	Fully Supervised (100%)	56.18	64.69	84.10	54.56	16.38	67.31	36.95	65.27	53.58	35.76	82.70	65.08
CenterPoint	Sparse Boxes (2%)	8.09	25.77	24.62	2.84	0.00	15.66	0.00	4.07	3.33	0.29	25.11	4.96
CoIn		12.47	33.79	38.70	6.85	0.00	20.67	7.81	11.51	2.85	3.36	34.85	8.50
HINTED		32.62	45.76	66.63	32.71	7.59	54.56	11.42	21.16	29.12	19.08	57.22	26.63
OC3D++	Sparse Coarse Clicks (0.2%)	24.88	38.12	56.37	23.47	1.10	33.30	3.43	16.18	24.67	8.81	54.42	27.00
OC3D	Coarse Clicks (1%)	44.04	49.87	78.91	46.18	8.01	50.34	27.68	60.70	28.80	10.70	74.06	54.98

Table 2: The multi-class results on the nuScenes val set. 'C.V.', 'Ped.', and 'T.C.' are short for construction vehicle, pedestrian, and traffic cones, respectively.

Comparison with State-of-the-art

Validation on KITTI. In Tab. 1, we conducted experiments to compare our approach with state-of-the-art label-efficient methods on KITTI. Following the mainstream approaches (Liu et al. 2022; Xia et al. 2023, 2024), we also adopted Voxel-RCNN (Deng et al. 2021) as the base detector. For coarse click annotations, the labeling time per instance is approximately 1.2 seconds, which is about 2 times faster than center-click labeling and 100 times faster than bounding box labeling. Despite employing a more lightweight annotation form, retaining only coarse click annotation, our OC3D still achieves comparable performance with other methods. In the Car-3D detection task, at the 'Easy' and 'Mod' difficulty levels, the gap between our performance and the previous best is only 3.3% and 0.2%, respectively. Meanwhile, at the 'Hard' difficulty level, our method achieved the best performance. In the Car-BEV detection task, at the 'Easy' difficulty level, the gap between our performance and the previous best is only 2.3%. Meanwhile, at the 'Hard' difficulty level, our method achieved the best performance. Especially, under conditions of an extremely cost-effective sparse clicking, the proposed OC3D++ demonstrates the capability to sustain consistent performance.

Validation on NuScenes. We also conducted validation experiments on the nuScenes dataset. To ensure a fair comparison, we follow the previous methods (Xia et al. 2023, 2024) to select the CenterPoint (Yin, Zhou, and Krähenbühl 2021) as the base detector. In Tab. 2, where OC3D achieves performance comparable to the fully supervised baseline method. This confirms that our SC3D still demonstrates outstanding performance on highly challenging multi-class tasks. In the experimental with sparse coarse click annotations, the mAP of OC3D++ decreased. The reason is the higher number of instances in the nuScenes, which poses a greater challenge for sparse settings.

Conclusion

We designed an efficient annotation strategy called coarse click annotation for 3D object detection and proposed a weakly-supervised object detection method, OC3D, leveraging this approach. In the mixed pseudo-label generation stage, we propose a novel method for dynamic and static classification, and design the Click2Box and Click2Mask modules according to the motion states of objects to generate mixed pseudo-labels. Subsequently, in the Mask2Box enhanced mixed supervised training stage, We train the 3D detector with mixed pseudo-labels and design the Mask2Box module to obtain richer supervisory information. Extensive experiments on KITTI and nuScenes have shown that our OC3D achieves commendable performance with purely clicks.

Limitations. The mixed pseudo-label generation stage provides the detector with initial labels describing instance geometry and location, but these rule-based labels lack the quality of human annotations.

References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, X.; Shi, S.; Zhu, B.; Cheung, K. C.; Xu, H.; and Li, H. 2022. Mppnet: Multi-frame feature intertwining with proxy points for 3d temporal object detection. In *European Conference on Computer Vision*, 680–697. Springer.

Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; and Jia, J. 2023. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21674–21683.

Deng, J.; Shi, S.; Li, P.; gang Zhou, W.; Zhang, Y.; and Li, H. 2021. Voxel R-CNN: Towards High Performance Voxel-based 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial Databases with Noise. *Knowledge Discovery and Data Mining,Knowledge Discovery and Data Mining*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 3354–3361.

Himmelsbach, M.; Hundelshausen, F. v.; and Wuensche, H.-J. 2010. Fast segmentation of 3D point clouds for ground vehicles. In *2010 IEEE Intelligent Vehicles Symposium*.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 12697–12705.

Liu, C.; Gao, C.; Liu, F.; Liu, J.; Meng, D.; and Gao, X. 2022. SS3D: Sparsely-Supervised 3D Object Detection From Point Cloud. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 8428–8437.

Meng, Q.; Wang, W.; Zhou, T.; Shen, J.; Jia, Y.; and Van Gool, L. 2021. Towards a weakly supervised framework for 3d point cloud object detection and annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8): 4454–4468.

Shi, S.; Jiang, L.; Deng, J.; Wang, Z.; Guo, C.; Shi, J.; Wang, X.; and Li, H. 2023. PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection. *International Journal of Computer Vision*, 131(2): 531–551.

Shi, S.; Wang, X.; and Li, H. 2019. PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 770–779.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; and et, a. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2443 – 2451.

Wang, H.; Cong, Y.; Litany, O.; Gao, Y.; and Guibas, L. J. 2021. 3DIoUMatch: Leveraging IoU Prediction for Semi-Supervised 3D Object Detection. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 14610–14619.

Wang, Y.; Deng, J.; Hou, Y.; Li, Y.; Zhang, Y.; Ji, J.; Ouyang, W.; and Zhang, Y. 2024. CluB: Cluster Meets BEV for LiDAR-Based 3D Object Detection. *Advances in Neural Information Processing Systems*, 36.

Wu, H.; Wen, C.; Li, W.; Yang, R.; and Wang, C. 2022. Learning Transformation-Equivariant Features for 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wu, H.; Wen, C.; Shi, S.; Li, X.; and Wang, C. 2023. Virtual Sparse Convolution for Multimodal 3D Object Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 21653–21662.

Xia, Q.; Deng, J.; Wen, C.; Wu, H.; Shi, S.; Li, X.; and Wang, C. 2023. CoIn: Contrastive Instance Feature Mining for Outdoor 3D Object Detection with Very Limited Annotations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6254–6263.

Xia, Q.; Ye, W.; Wu, H.; Zhao, S.; Xing, L.; Huang, X.; Deng, J.; Li, X.; Wen, C.; and Wang, C. 2024. HINTED: Hard Instance Enhanced Detector with Mixed-Density Feature Fusion for Sparsely-Supervised 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15321–15330.

Yan, Y.; Mao, Y.; and Li, B. 2018. SECOND: Sparsely Embedded Convolutional Detection. *Sensors*, 18.

Yang, Y.; Fan, L.; and Zhang, Z. 2024. MixSup: Mixed-grained Supervision for Label-efficient LiDAR-based 3D Object Detection. In *ICLR*.

Yin, T.; Zhou, X.; and Krähenbühl, P. 2021. Center-based 3D Object Detection and Tracking. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, D.; Liang, D.; Zou, Z.; Li, J.; Ye, X.; Liu, Z.; Tan, X.; and Bai, X. 2023. A Simple Vision Transformer for Weakly Semi-supervised 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8373–8383.

Zhang, X.; Xu, W.; Dong, C.; and Dolan, J. M. 2017. Efficient L-shape fitting for vehicle detection using laser scanners. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, 54–59.